Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 73 (2015) 48 – 55

Procedia Computer Science

The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

A Tutorial on Speech Synthesis Models

Tabet Youcef [a]*, Boughazi Mohamed [b], Affifi Sadek [b]

[a]University of Boumerdes,35000 Algeria
[b]University of Annaba,23000 Algeria

Abstract

For Speech Synthesis, the understanding of the physical and mathematical models of speech is essential. Hence, Speech Modeling is a large field, and is well documented in literature. The aim in this paper is to provide a background review of several speech models used in speech synthesis, specifically the Source Filter Model, Linear Prediction Model, Sinusoidal Model, and Harmonic/Noise Model. The most important models of speech signals will be described starting from the earlier ones up until the last ones, in order to highlight major improvements over these models. It would be desirable a parametric model of speech, that is relatively simple, flexible, high quality, and robust in re-synthesis. Emphasis will be given in Harmonic / Noise Model, since it seems to be more promising and robust model of speech.

Peer-review under responsibility of organizing committee of the International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

Keywords: Linear Prediction Model; Sinusoidal Model; Harmonic/Noise Model

* tabet2402@yahoo.fr

Peer-review under responsibility of organizing committee of the International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)
doi:10.1016/j.procs.2015.12.048

1. Introduction

Speech is the most natural form of human communication. The temporal-spectral variations of speech signals convey such information as words, intention, expression, intonation, accent, speaker identity, gender, style of speaking, state of health of the speaker and emotion[1].

The time signal evolution can either be represented by a model or not. The advantages of using a model are its capacity to reduce the acoustic signal's redundancy and to define parameters that are better fitted to acoustic processing[2].

The speech production model and the sinusoidal model are the two main models used in speech synthesis. The first one is a model with several in series-system that represent the different stages of the human speech production, i.e. excitation system, vocal tract, and lips system. The second model decomposes the observed signal into a sum of sinewave components, i.e. a sum of frequency and/or amplitude modulated cosines. Selecting a model depends on many factors such as quality of synthesized speech, ease of parameter extraction, modification of parameters, number of parameters and computation load.

The remainder of this paper is organized as follows. First, related works is given in section 2. Next, in section 3 we present the source filter model. After that the linear prediction model is depicted in section 4. In section 5, the sinusoidal model is described. This is followed by the harmonic/noise model in section 6. Finally, conclusions are given in section 7.

2. Related works

Several approaches have been proposed for speech modeling, some of them will be briefly discussed below and the most important one will be described in detail in the next sections.

Serra[3] suggested a hybrid system for the analysis, transformation, and synthesis of sound based on a deterministic/stochastic decomposition. This system is designed to obtain musically useful intermediate representations for sound transformations. The deterministic component is represented by a series of non-necessary harmonic sinusoids calculated by Short Term Fourier Transform (STFT)-based peak-picking method. The stochastic component is represented by a series of magnitude-spectrum envelopes that work as a time varying filter excited by a white noise. This approach is able to create new sounds out of the representation of a particular sound. The deterministic signal is obtained by synthesizing a sinusoid from each trajectory. Then, the residual between the deterministic component and the original sound is modeled by a series of envelopes. Finally, the stochastic signal is generated by an inverse STFT. This system is very flexible and allows for transformations of the sound by manipulating each component separately.

George and Smith[4], in their work, used a system based on the combination of an overlap-add (OLA) sinusoidal model with an analysis-by-synthesis (ABS) technique to determine the sinusoidal model parameters. They introduce an equivalent frequency domain algorithm. In addition, a refined overlap-add sinusoidal model is derived. There is a well correlation between the refined overlap-add synthesis and analysis-by-synthesis. The proposed analysis-by-synthesis/Overlap-add (ABS) system achieves very high synthetic speech quality.

To model the *transient* part of speech signal, several models have been proposed. For example, a flexible analysis/synthesis model for transient signals is proposed in[5]. The model presented is a parametric model for transients that allows for a wide range of signal transformations. In[6] a model that gathers the properties of a sinusoidal representation and an OLA processing step is presented. This model is acknowledged as being efficient for rendering of time-localized events. Also, the authors in [7, 8, 9] proposed the Exponentially Damped Sinusoidal Model (EDSM), along with more powerful parameter estimation schemes based on either matching pursuit or subspace methods. Subspace methods have good spectral properties and do not suffer from the time frequency trade-off embedded in other methods. However, they are computationally intensive.

Recently, in order to refine the proposed models using more powerful parameter estimation schemes, Fan-chirp transform that employs an adaptive analysis basis composed of quadratic chirps is presented in[10]. A sinusoidal analysis of speech similar to the model proposed in [11] but with the Fan-chirp transform instead of the Fast Fourier Transform has been conducted [12,13] with very satisfactory results.

3. Source Filter Model

In the *Source-Filter* decomposition, the speech signal results from the combination of some acoustic energy (interaction of the lungs and the larynx) coupled with a transfer function that is determined by the shape of the supra-glotic cavities. In the context of signal processing, the Source-Filter model describes the speech signal as the convolution of an excitation signal by a time varying filter. The excitation characterizes the variation of acoustic pressure in the larynx and the filter represents the time-frequency behavior of the vocal tract transfer function[14].

For the most part, it is sufficient to model the production of a sampled speech signal by discrete-time system model. In this model the unvoiced excitation is assumed to be a random noise sequence, and the voiced excitation is assumed to be a periodic impulsion train with impulses spaced by the pitch period rounded to the nearest sample[15]. The discrete-time system model is also called *source filter* model. This system can be described by the convolution expression

$$s(n) = \sum_{m=0}^{\infty} h(m)u(n-m) \tag{1}$$

Where s(n) is the entire speech signal, h(n) is the impulse response, and u(n) is the voiced/unvoiced excitation.

To simplify analysis, it is often assumed the system is an all-poll filter with system function of the form

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

where G and p are respectively the gain and the order of the filter.

The linear system is assumed to model the composite spectrum effects of radiation, vocal tract tube, and glottal excitation pulse shape (for voiced speech only). Over a short time interval the linear system in the model is commonly referred to as simply the “vocal tract system” and the corresponding response is called the “vocal tract impulse response”.

For all-pole linear systems, as represented by (2), the input and the output are related by a difference equation of the form

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{3}$$

This *Source filter* model is intimately related to the *Linear Prediction* model discussed in the following section.

4. Linear Prediction Model

In 1960, Fan introduced a *linear model* of the time-domain waveform of the speech signal. To the Source-Filter hypothesis, he added the hypothesis of independence between the glottal waveform and the vocal tract. The vocal tract is modeled as an all pole filter, also called “Auto Regressive”. The glottal waveform is modeled roughly as a pulse train with a fundamental period equal to the pitch (for voiced sound) and as a white noise with zero mean and unit variance (for unvoiced sounds)[2, 14].

The term “Linear prediction” refers to the mechanism of using a linear combination of the past time-domain samples, s(n-1),s(n-2),…, s(n-p), to approximate or to predict the current time-domain sample s(n)[16,17]. So, a linear predictor of order P, with prediction coefficients $\{\alpha_k\}$, is defined as a system whose output is

$$s_p(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{4}$$

Where $s_p(n)$ is the predicted signal
The prediction error is given by

$$e(n) = s(n) - s_p(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{5}$$

Equation 5 can be presented in the z-domain as

$$E(z) = s(z)A(z) \tag{6}$$

Where E(z) is the z-transform of e(n), S(z) is the z-transform of s(n), and A(z) is the z-transform of the prediction error filter given by

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{7}$$

Upon further inspection of equation 3 and 5, it can be seen that if the model is exactly accurate for the speech signal, and if $\{a_k\} = \{\alpha_k\}$, then e(n) = Gu(n).
Thus, A(z) becomes the inverse filter of the system H(z) of equation 2

$$H(z) = \frac{G}{A(z)} \tag{8}$$

The main problem of linear predictive analysis thus becomes the estimation of the predictor coefficients $\{\alpha_k\}$ so that the prediction error e[n] is minimized under some criterion. The mean squared error is by far the most utilized optimization criterion. The coefficients $\{\alpha_k\}$ that minimize the mean squared error are assumed to be the parameters of the system function H(z) of equation 2.
The squared prediction error E_n in a short-time frame $s_n(m)$ starting at sample n is defined as

$$E_n = \sum_m e_n^2(m) \tag{9}$$

$$E_n = \sum_m (s_n(m) - s_{pn}(m))^2 \tag{10}$$

$$E_n = \sum_m (s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k))^2 \tag{11}$$

Where $s_n(m) = s(n+m)$

Two major approaches to the computation of the LPC coefficients have been developed: the autocorrelation method and the covariance method. Hence, the minimization of the equation 11 leads to normal equations which can be solved using several algorithms.

The Linear predictive analysis is applied on a frame-by-frame basis to the speech signal. Hence, for each frame a linear predictive filter is generated. This filter models the glottal excitation pulse shape, vocal tract and lip radiations effects. During voiced speech, a simple pulse train excites the linear predictive filter. However, for unvoiced speech the filter is excited by a white noise.

The method of linear predictive analysis was one of the most powerful speech analysis techniques because it is simple, fast, and has a limited number of parameters. The main drawback of this method is that is inherently "buzzy" due to its parametric nature, and this degrades the speech quality. Also, phonemes such nasals cannot be modeled by the linear prediction model because they contain anti-formants, and this model is an all-pole model. However, the method of linear prediction has been the predominant technique for estimating the basic speech parameters, e.g., pitch, formants, spectre, vocal tract area functions, and for representing speech for low bit rate transmission or storage until the end of the 1980s, after which it gave way to more complex techniques which offered a better signal quality, e.g., sinusoidal model and its derivatives.

5. Sinusoidal Model

The *Sinusoidal model* presented in [11], represents a speech as a sum of sinusoidal functions, evolving over time. This model has the capacity of frequency resolution (sinus resolution) and time resolution (evolution of each sinusoid over time). The speech signal is assumed to be the output of a slowly time varying digital filter with an excitation that capture the nature of the voiced/unvoiced distinction in speech production (Excitation expressed as a sum of sinusoids). The speech signal resulting from the full model is written

$$s(n) = \sum_{l=1}^{L} a_l(n) \cos(w_l n + \theta_l) \tag{12}$$

Where L is the number of sinusoidal segments, a, w, θ represent respectively the amplitude, frequency and phase of the sinewave.

The *analysis* scheme is based on the notion of sinusoidal track, referring to the components of the sum in the synthesis formula 12. The number L of tracks varies with time: each track is active during a given lapse of time and this has to be determined by a tracking algorithm. It is thus necessary to estimate the number of components, their amplitudes, and frequencies. This analysis step is based on the Short Term Fourier Transform. For each frame, the spectral peaks are obtained by searching for all local maxima on the amplitude spectrum and then eliminating those whose amplitude is below a given threshold. The position of the peaks provides frequencies and amplitudes of the sinusoidal components. Phases of these components are calculated as the phase of the Short Term Fourier Transform for a given frequency. For each frame, a set of L spectral peaks is thus obtained.

The *synthesis* signal is calculated by an overlap-add of the short-term signal from equation 12. In this case, the sinusoidal tracks are not explicit. It is much more advantageous to follow the sinusoidal tracks explicitly and then to interpolate the synthesis parameters along these tracks. So, this alternative synthesis method is performed in several steps. The first one is the parameter matching, the second one is the parameter interpolation, and finally the synthetic waveform is calculated from equation 12.

The matching procedure is done between parameter values computed at two consecutive frame boundaries in order to solve the problem of the variable number of the peaks across frames. The algorithm proposed in[11] has to be put into operation for detecting the "birth", "continuation", and "death" of the sinusoidal components across frames. After applying parameter matching algorithm, the next step is about parameter interpolation in order to avoid abrupt changes from one frame to the next. The amplitudes are linearly interpolated between two successive frames and the phases and frequencies are interpolated using a cubic function.

Finally, equation 12 is used to calculate the synthetic speech signal.

The main advantage of the sinusoidal model is that it performs speech modification by finding the sinusoidal components for a waveform and performing modification by altering the parameters of the equation 12, namely the amplitudes, phases, and frequencies[18]. So, the speech delivered by this model is perceptually identical to the original

one. Because, the sinusoidal model is suited for modeling harmonic sounds, the unvoiced sounds are poorly represented by this model. Also, this model has limitation like large database and computational complexity.

The need of more complex models that could handle the non-harmonic component of speech sound led to the Deterministic/Stochastic model[3] or to the Harmonic/Noise model[19] .

6. Harmonic/Noise Model

Harmonic/Noise Model (HNM), divides the speech signal into two parts: the harmonic part and the noise part. The harmonic one represents the quasi-periodic components of the speech signal such as vowels and some voiced consonants and the noise part represents the non-periodic part such as fricative aspiration noise, bursts, unvoiced speech, etc. The harmonic part is modeled through a set of harmonically related sinusoids with slowly varying amplitudes and frequencies. However, the noise part is usually modeled as white Gaussian noise passing through a shaping filter. The speech spectrum is divided into two sub bands delimited by a time varying maximum voiced frequency.

A number of models based on the principle of decomposing the speech signal into harmonic part and noise part have been proposed. Given its popularity we explore here the Harmonic/Noise model proposed in[19, 20, 21].

The speech signal in such model can be expressed by a combination of harmonic and noise like models as

$$s(n) = s_h(n) + s_n(n) \tag{13}$$

The signal in the harmonic part can be modeled as

$$s_h(n) = \sum_{l=1}^{L} A_l(n)\cos(l\theta(n) + \phi_l(n)) \tag{14}$$

where

$$\theta(n) = \int_{-\infty}^{n} \omega_0(u)du \tag{15}$$

And where $A_l(n), \phi_l(n)$ denote the amplitude and phase at time n of the kth harmonic respectively, ω_0 is the fundamental frequency and L is the number of harmonics included in the harmonic part.

However, the noise part can be modeled as

$$s_n(n) = e(n)[h(n;\tau) * b(n)] \tag{16}$$

where b is a white Gaussian noise; h is a time varying normalized all-pole filter; e is an energy envelope function applied to give the filtered noise the correct temporal pattern.

Analysis: The estimation of the pitch is the first step in Harmonic/Noise Model Analysis. It is obtained by using a time domain approach, i.e. searching the minimum value of an error function (an autocorrelation approach). From this initial pitch estimation, a harmonic model is fitted to each frame and the voiced/unvoiced decision is made by using criterion which takes account into how close this harmonic model is to the original model.

For voiced frames, we then estimate the maximum voiced frequency Fm. This estimation is based on a peak picking algorithm. Once the maximum voiced frequency has been found, accurate pitch estimation is necessary. This is done by minimizing a given error between the initial pitch estimation and the range of frequencies classified as voiced from the previous step. The amplitudes and phases of the harmonics are found in time domain using a weighted least square error between the real and the synthetic wave form.

For the estimation of the parameters of the noise component (unvoiced part), in each analysis frame, a spectral density of the original signal is modeled by an autoregressive filter. This filter will be excited by a white noise and

the dynamic characteristics are considered by using a variance envelope which modulates the excitation. Also, a triangular- like time domain energy envelope modulates the noise comprising the second part of a voiced spectrum. A high pass filter of F_m (maximum voiced frequency) cut frequency is used to separate the harmonic part from the noise one.

The final step in Harmonic/Noise model analysis is phase modification for speech synchronization. This is done by using a time domain technique to adjust the relative positions of the waveforms within the frames to ensure that the all align.

Synthesis: The synthesis is performed in a pitch synchronous way. The amplitudes and the phases of the harmonic component, are estimated via a least square criterion, and are linearly interpolated between successive frames. Only, the phases are unwrapped before applying interpolation.

An overlap-add (OLA) method is used to synthesis the noise part: A Gaussian noise b is passed through the filter h several times per frame in order to ensure that the temporal characteristics are successfully generated; the noise is high pass filtered with a cut-off frequency equal to the maximum voiced frequency Fm. Next, to ensure that the noise is synchronized with the harmonic part, it is modulated in the time domain.

The final synthetic speech signal is obtained by adding the two parts, i.e. harmonic part plus noise part.

Investigations show that Harmonic/Noise model outperforms almost all models of speech signal in term of naturalness, indelibility and pleasantness which are of pre-requisite in many speech synthesizers[20,21].

Harmonic/Noise model is a pitch synchronous system, and unlike other concatenative approaches hence it eliminates the problem of synchronization of speech frames, and shows the capabilities of providing high quality prosodic modifications without business when compared to other methods[22].

For these reasons Harmonic/Noise model was chosen by AT&T to serve as the backend in their Next Generation Text to Speech (TTS) System which is claimed to produce extremely high quality synthetic speech. Also Harmonic/Noise model is widely used in other frameworks. For example, it has been quite naturally involved in the development of voice modification, transformation, and conversion systems[23, 24, 25]. Harmonic/Noise model has been used for the development of a high quality vocoder applicable in statistical framework particularly in modern speech synthesizer[26].

One main drawback of the harmonic noise model is its complexity[19].

7. Conclusions

In this paper, we have presented an overview of several speech models used in speech synthesis.

Several approaches have been proposed for speech synthesis modeling, some of them have been briefly discussed and the most important ones, namely, Linear Prediction model, sinusoidal model, and Harmonic/noise model were discussed in some details.

The linear prediction model was one of the most important speech synthesis models until the end of the 1980s, after which it gave way to more complex techniques which offered a better signal quality, e.g., sinusoidal model and its derivatives.

Sinusoidal model is especially suited for modeling voiced sounds, due to the incapability of the model to capture noisy sound well, it is necessary to use other types of models which decompose the speech signal into several parts as the deterministic/stochastic model, or the Harmonic/Noise model. However, these models cannot be used to model the transient part of speech signal. The solution is to use another type of models called Harmonic/Transient/Noise model or Deterministic/Transient/Stochastic model.

Compared to the existing speech synthesis models, Harmonic/Noise model shows more practical potential for speech synthesis. It is widely used in several speech synthesis frameworks as it offers high quality speech with a relatively smaller number of parameters, and with ease pitch and time scale modification.

To improve the quality of the synthesized speech, new methods for extracting certain parameters of the Harmonic/Noise Model are proposed in the literature.

References

1. Saeed V. Vaseghi. Multimedia Signal Processing.Theory and Applications in Speech, Music and Communications.Wiley. 2007
2. MarianiJ.Language and speech processing. ITSE, WELLEY. 2009.
3. Serra X. A System for Sound Analysis, Transformation, Synthsis based on a Determistic plus Stochastic Decomposition.PhD thesis, Stanford University.1989.
4. George E.B.andSmith M. J .T. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model.IEEE Transactions on Speech and Audio Processing.1997. 5(5):389– 406.
5. Verma T., Levine S., and Meng T. Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals. In Proc. of the ICMC 1997, Thessaloniki, Greece. 1997. p. 164–167.
6. Peeters G. and RODET X. Sinola. A new analysis/synthesis method using spectrum peak shape distortions, phase and reassigned spectrum.In Proceedings of the International Computer Music Conference (ICMC'99).1999.
7. Nieuwenhuijse J., Heusdens R., and Deprettere E. Robust exponential modeling of audio signals. InICASSP 1998, Seattle, WA, USA.May 1998.
8. Jensen J, Jensen S. H, and Hansen E. Exponential sinusoidal modeling of transitional speech segments.In Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP). 1999. Vol 1, p. 473–476.
9. Jensen J., Heusdens R., andJensen S. A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids.IEEE Transaction on Speech and Audio Processing. 2004. 12(2):121–132.
10. Kepesi M. and Weruaga L. Adaptive chirp-based time-frequency analysis of speech. Speech Communication,2006. 48:474–492.
11. Mcaulay R.J., Quatieri T.F. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. on ASSP.1986.vol. 34, no. 4.
12. Dunn R. B. and Quatieri T. F. Sinewave Analysis/Synthesis Based on the Fan-Chirp Transform. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). October, 2007.
13. Dunn R. B., Quatieri T. F., and MalyskaN. Sinewave parameter estimation using the fast fan-chirp transform.InWorkshop onApplications of Signal Processing to Audio and Acoustics (WASPAA). 2009. p. 349–352.
14. Fant G. Acoustic Theory of Speech Production, Mouton, La Hague-Paris. 1960.
15. Rabiner L. R. and Schafer R. W. Introduction to digital speech processing. Foundations and Trends in Signal Processing. 2007. vol. 1, no. 1, p. 1-194.
16. Rabiner L. R. and Schafer R. W., Digital Processing of Speech Signals. Prentice-Hall Inc. 1978
17. Atal B. S. and Hanauer S. L. Speech analysis and synthesis by linear prediction of the speech wave. Journal of the Acoustical Society of America. 1971. p. 637-655.
18. Taylor P. Text-to-Speech Synthesis.Cambridge University Press. 2009
19. Stylianou Y. Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification. PhD thesis, E.N.S.T – Paris. 1996.
20. Laroche J., Stylianou Y., and Moulines E. HNS: Speech modification based on a harmonic + noise model. Proceedings of the International Conference on Acoustics, Speech, and Siqnal Processing. 1993. 2:550-553.
21. Stylianou Y., Laroche J., and Moulines E. High quality speech modification based on a harmonic + noise model. Proceedings of EURO SPEECH . 1995. p. 451-454.
22. Syrdal A., Stylianou Y., Garrison L., Conkie A., and Schroeter J. TDPSOLA versus harmonic plus noise model in diphone based speech synthesis. Proceedings of the International Conference on Acoustics, Speech and Signal Processing.1998. p. 273-276.
23. Stylianou Y., Cappé O. and Moulines E. Statistical methods for voice quality transformation.In Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'95).1995.
24. Stylianou Y. Cappé O. and Moulines E. Continuous probabilistic transform for voice conversion.IEEE transactions on Acoustics, Speech and Signal Processing (ASSP).1998. 6(2):131-142.
25. Oudot M. Application du modèle sinusoïdes et bruit au décodage, au débruitage et à la modification des sons de parole. PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris. 1998.
26. Sainz I., Navas E., Hernaez I. Harmonic plus noise model based vocoder for statistical parametric speech synthesis. Selected Topics in Signal Processing, IEEE. 2014. 8(2).